DATA 319: Model-based and Data-based Methods for Data Analytics

Fall 2024

Course Logistics

Prefix/Number:	DATA 319.1 (WSU Everett); DATA 319.1 (WSU Pullman); DATA 319.1 (WSU Vancouver)
Credit Hours:	3
Prerequisites:	• DATA 219, CPT S 215, CPT S 223, or CPT S 233 *and* • MATH 220 or MATH/DATA 225 *and* • STAT 360
Classroom:	Room 340 (WSU Everett); SLOA 7 (WSU Pullman); VECS 125 (WSU Vancouver)
Class Time:	M/W/F 10:10-11:00 am
Final Project:	*The time and the place of the presentation will be determined later during the semester*

Instructor Information

Instructor:	Gani Nurmukhametov
Office:	WSU Everett Room 404
Phone:	(425) 405-1659
Email:	gani.nurmukhametov@wsu.edu
Office Hours:	typically[1] M 11 am - 12 noon, F 9-10 am or by prior appointment in Zoom

Introductory Note

First of all, welcome to DATA 319! I am looking forward to meeting you and helping out with your journey at Washington State University.

Secondly, sorry this is such a long document. Despite that, I hope you will make time to read it through at least once during the first week of the course as it contains a useful summary of the material we will be covering and will serve as a repository of important information and links.

Finally, definitely feel free to reach out if you have any questions or concerns about the course or material. I am always happy to chat about data, statistics, or academic life and would welcome opportunities to offer my perspective or simply serve as a sounding board.

Course Description

This course provides an introduction to modern modeling methods for data analysis, focused on applications to high dimensional data. As datasets with hundreds or thousands of variables have become more common and easier to manipulate, an understanding of the relevant theoretical and practical concerns is an increasingly important part of data analytics.

This focus on the nuances of high-dimensional data will allow us to explore the properties of many real-world data sets and gain hands-on experience with analytical methods implemented in the Python programming language. Throughout the course we will also build on students' foundational knowledge of mathematics and statistics to introduce multivariate concepts. The skills learned in this course are relevant to careers in industries specializing in large multivariate data sets, including (but not limited to!) data analytics, operations research, biological sciences and engineering.

Expectations for Student Effort

You are expected to spend a minimum of 9 hours per week for a three-credit course, of which 3 hours are spent on instructor-led activities (lectures and discussions) and 6 hours are spent on outside activities. These outside activities include, but are not limited to: reading, studying, problem solving, writing, homework, and other preparations for the course. Achievement of course goals may require more than the minimum time commitment. For the most accurate and up to date information go to Academic Regulations.

[1] I may change the time or even cancel office hours on a given day. Please follow my announcements on Canvas.

Course Materials

We will use these three textbooks:

1. Applied Multivariate Statistical Analysis by Härdle and Simar; hereinafter referred to as *HS*.
2. Mining of Massive Datasets by Leskovec, Rajaraman, and Ullman; hereinafter referred to as *MMDS*.
3. An Introduction to Statistical Learning with Applications in Python by James, Witten, Hastie, Tibshirani, and Taylor; hereinafter referred to as *ISLP*.

You may also find these textbooks useful:

- Applied Multivariate Statistical Analysis by Johnson and Wichern
- Numerical Algorithms by Solomon
- Linear Algebra and Its Applications by Lay, Lay, and McDonald
- Introduction to Linear Algebra by Strang
- Linear Algebra and Learning from Data by Strang

Software

Each course module incorporates computational examples working in the Python programming language. Prior Python experience at the level of DATA 219 or 302 will be assumed. These sections will reinforce the main material by allowing you to perform and implement the analyses discussed in the study materials and videos. Some homework problems and all of the projects will also require the use of computational methods and previous programming experience and knowledge of statistical software tools from prerequisite courses will be assumed. Throughout the course we will help students build familiarity and experience with standard analytics libraries including pandas, numpy, matplotlib, scikit-learn, and seaborn. Following resources on Python might be useful:

- Python Data Science Handbook by VanderPlas (free resource)
- Data Science from Scratch: First Principles with Python by Grus, 2nd edition.
- Learning Data Science: Data Wrangling, Exploration, Visualization, and Modeling with Python by Lau, Gonzales, and Nolan.
- Murach's Python for Data Science by McCoy, 2nd edition.

Some problem set questions are completed on Google's Colaboratory. You are required to sign in using your GoogleDrive credentials to work on Google Colab. If you do not feel comfortable sharing your personal email credentials, then consider creating a unique google email address for this class.

Class Communication and Instructor Interaction

Announcements and other official communications will be posted on the Canvas website as well as sent to your official WSU email accounts. You should check these messages regularly to stay informed about upcoming due dates and updates to the syllabus.

I am accessible by email at gani.nurmukhametov@wsu.edu. I teach multiple courses, so please, as a courtesy to me, include "DATA 319 (Fall 2024, Synch)" in the subject line for any messages concerning this course. I will most likely read your email as soon as I receive it, but it still may take me a little while to respond. I will try to get back to you as quickly as I can but this does mean that queries sent immediately before a deadline may not receive substantive responses in time to be directly helpful, so please plan ahead ☺

I strongly encourage you to utilize the resource that is freely available to you, that is my office hours! This will allow you to get immediate feedback from me. I hold my office hours in Zoom. Following the Zoom meeting link, you will appear in the waiting room first, and if I am not already talking to another student, then I will accept you right away. If you cannot attend my office hours, do not hesitate to send me an email and we can schedule a short appointment in Zoom that fits both you and me.

Student Learning Outcomes

Students who successfully complete the course will be able to:

- Construct scripts for processing, analyzing, and visualizing high-dimensional data in Python
  - **Relevant Assessments:** All Problem Sets, Quizzes, Final Project
- Conduct exploratory multivariate data analysis, including constructing relevant visualizations
  - **Relevant Assessments:** Problem Set 1, Midterm Exam
- Apply and analyze the MapReduce framework and implement simple tasks using PySpark on Colab
  - **Relevant Assessments:** Problem Set 1, Problem Set 2
- Compute matrix factorizations and interpret the resulting outputs in the context of the original data
  - **Relevant Assessments:** Problem Set 1, Midterm Exam
- Use frequent itemset modeling and association rules to characterize relevant data
  - **Relevant Assessments:** Problem Set 2, Midterm Exam
- Apply the A Priori algorithm and its variants to frequent itemset problems
  - **Relevant Assessments:** Problem Set 2, Midterm Exam
- Understand and implement high-dimensional data embeddings in $\mathbb{R}^n$
  - **Relevant Assessments:** Problem Set 2, Problem Set 4
- Select and apply relevant metrics for (dis)similarity on high-dimensional data
  - **Relevant Assessments:** Problem Set 2, Final Project
- Select and perform appropriate methods for dimension reduction, including PCA, MDS, Factor Analysis, and Discriminant Analysis
  - **Relevant Assessments:** Problem Set 2, Problem Set 4
- Define the characteristics and properties of the multivariate normal distribution
  - **Relevant Assessments:** Problem Set 3, Midterm Exam
- Characterize the concept of interdependence among multidimensional data through correlation
  - **Relevant Assessments:** Problem Set 3, Midterm Exam
- Implement appropriate multivariate null hypothesis testing
  - **Relevant Assessments:** Problem Set 3, Midterm Exam
- Perform, justify, and interpret the use of unsupervised analytic methods including clustering
  - **Relevant Assessments:** Problem Set 5, Final Project
- Perform, justify, and interpret the use of supervised analytic methods
  - **Relevant Assessments:** Problem Set 6, Final Project
- Identify ethical concerns relevant to designing and implementing data models and analyze potential impacts using appropriate frameworks
  - **Relevant Assessments:** Reading and Discussion Assignments, Final Project

Attendance and Participation

Due to the nature of the course (with students joining synchronously from several WSU campuses) and the wide range of topics that we will cover, daily attendance will be essential for your success. Although it is not officially a part of the course grade, missing class could adversely affect your grade by impacting your understanding of the material.

While I do not take an attendance roll call, 10% of your course grade comes from the in-class discussions. These discussions are based on the short readings about a relevant data analytic topic. However, if you miss a class on a day of the in-class discussion of if you did not participate in the discussion in class, then it is your responsibility to leave a detailed comment on the Canvas discussion board within one hour after the end of the class to get the credit for the assignment.

That said, I understand that you may occasionally have difficulties keeping up with the pace of the course. Please communicate with me in advance if possible, so I can help point you to useful resources.

Assignments and Assessments

The total number of points you can get in this course is 100 (plus up to 2 extra credit points), so you can treat points from the course assignments and assessments as percentages of your overall course grade.

- **Reading and Discussion Assignments (RDA):** These assignments involve completing several short readings on a relevant data analytics topic. A set of discussion questions related to the reading will be provided and you will be required to make a discussion post responding to those questions. These are more subjectively graded and will receive a score - the main goal here is participation. There will be 10 RDAs in total, each worth 1 point. Late submissions are allowed but they are subject to a possible late penalty at my discretion.

- **Problem Sets:** These assignments will usually be a mixture of direct questions about the lecture material and opportunities for you to apply the methods we discuss to real data. There will be 6 problem sets in total, each worth 5 points, and the highest score will count twice towards the course grade. Submit your work on Canvas before the due date specified, late submissions will not receive credit (except for the prior accommodations). You are required to submit your work both as a completed Jupyter notebook file and as a .html file. If a problem set has a question using Google Colab, then save and submit your answers (both a Jupyter notebook and a .html file) separately from the other problem set questions. You will work on problem sets in the teams of 3-5 students. Only one submission from a team is expected; however, include only the names of the students who *actually contributed* to the team work on the current problem set. Students who failed to put work on a problem set will receive zero credit for that assignment.

- **Quizzes:** These individual assignments cover both theoretical concepts and some tasks performed in Python. There will be 5 quizzes in total, each worth 2.5 points, and the highest score will count twice towards the course grade. Quizzes are timed to be completed in 120 minutes or less.

- **Midterm Exam**: There will be a single exam at the end of Week 7, worth 15 points. This individual assignment is an open notes exam, and it will be submitted on Canvas. The exam is timed to be completed in 240 minutes or less.

- **Final Project:** The final assessment in the course is to complete a project employing the multivariate methods discussed throughout the class on a large dataset. You will work in the teams of 3-5 students on your final project (same teams as for the problem sets). The final project is worth 25 points. There are three deliverables regarding the project: a completed team contract (worth 1 point) is due the end of Week 2, a project proposal (worth 2 points) is due the end of Week 11, and a project report/presentation (worth 12 points) due the end of the Finals Week. More details will be provided later in the course.

Remaining 10 points will be assigned based on a peer evaluation, where you will anonymously evaluate the contribution of your teammates to a final project; 2 points will be given for filling in an evaluation form, and up to 8 points will be given based on the evaluation feedback from your teammates. Peer evaluation is due the end of the Finals Week.

- *Extra credit (optional):* There will be two optional surveys: an introduction survey is due the end of Week 1 and an exit survey is due the end of the Finals Week. Both surveys are worth 1 point each.

Grading Policy

The "weights" of course assignments towards the overall course grade is as follows:

Assignment	Percentage
RDA	10%
Problem Sets	35%
Quizzes	15%
Midterm Exam	15%
Final Project	25%
Extra credit (optional)	*2%*

I will determine your letter grade using the following grade schema:

A	93-100
A-	90-92.99
B+	87-89.99
B	83-86.99
B-	80-82.99
C+	77-79.99
C	73-76.99
C-	70-72.99
D+	65-69.99
D	60-64.99
F	0-59.99

Late Work Policy

Late submissions are not normally allowed (except for RDAs, where they are subject to a late penalty). Earlier submissions are allowed at any time before the due date. Extensions *may be* allowed by contacting me well in advance and in case of real emergencies. After the answers for the assignments have been provided on the Canvas website, that problem set or exam cannot be made up, so please reach out to me early if needed. It is better to turn in a partially completed homework than none at all.

For an excused missed midterm exam, the accommodation is at my sole discretion and may include a reweighting of the remaining components making up the student's grade or taking a makeup exam. Again, please reach out to me well in advance.

Collaboration Policy

You are encouraged (and sometimes required) to work with other students for the assignments in the class. However, the work that you submit for the *individual* assignments should be your own and in particular should be written in your own words and communicate your own understanding of the solution. If you do collaborate, please list the names of the other students you worked with on your submission. You may be asked to explain your work in person to obtain full credit. Obtaining solutions for course problems from external sources will be considered a violation of the academic integrity policy with consequences described below.

Online Discussion Policy

The essence of education is exposure to diverse viewpoints. In your discussion posts you'll meet students with vastly different opinions and backgrounds. You're encouraged to disagree with the substance of others' ideas and opinions but do so with an active sense of respect for one another, and without losing focus on the topic at hand. Personal attacks, inflammatory statements, flaming, trolling, and disruption of the discussion do not have a place in academic discourse. Postings must comply with University policy on use of computing resources, including those regarding harassment and discrimination, as well as conform to the WSU Community Standards.

I will aim to promote high-quality academic discussions by removing any posts I view as disruptive of the educational process and alerting students whose posts have been removed that they have violated course expectations. Students who continue to misuse the discussion boards after a warning may be subject to removal of access rights, course failure, and referral to the Office of Community Standards. Visit WSU Netiquette Guidelines.

Academic Integrity

Academic integrity is the cornerstone of higher education. As such, all members of the university community share responsibility for maintaining and promoting the principles of integrity in all activities, including academic integrity and honest scholarship. Academic integrity will be strongly enforced in this course.

You are responsible for reading WSU's Academic Integrity Policy, which is based on Washington State law. Students who violate the Academic Integrity Policy will fail the assignment, will not have the option to withdraw from the course pending an appeal, and will be reported to the Center for Community Standards. Multiple violations of the policy will cause you to fail the course.

Cheating includes, but is not limited to, plagiarism and unauthorized collaboration as defined in the Standards of Conduct for Students (identified in Washington Administrative Code (WAC) 504-26-010(2)). You need to read and understand all of the definitions of cheating. If you have any questions about what is and is not allowed in this course, please reach out to me before proceeding.

If you wish to appeal my decision relating to academic integrity, please use the form at the Center for Community Standards website. You must submit this request within 21 calendar days of the decision.

Incomplete Grade Policy (Academic Rule 90h)

Incompletes are granted only with my permission and are subject to the following guidelines:

1. You must request an incomplete in writing or by e-mail to me before the end of the semester. This request must be signed and dated by you (or identified by your WSU e-mail address) and must explain the reasons behind the request for the incomplete.
2. In order to be considered for an incomplete grade, these two conditions should be met:
    - You must complete a minimum of 75 percent of the assigned course work.
    - You must have a mathematical possibility of scoring a 60 percent or above for the entire course.
3. If extraordinary circumstances (e.g., family emergency, serious illness) are involved, I retain the discretion to grant an incomplete even if the minimum conditions outlined in item 2 above are not met.

If an incomplete grade is granted, the standard WSU policy applies (i.e., ALL work must be completed within one full year from the end of the enrollment semester at issue, unless a shorter time is specified by the instructor. Otherwise, an automatic grade of "F," or failing, will be entered on your transcript).

Library Support

All students enrolled in Washington State University online courses can use the WSU Libraries online databases and receive reference and research assistance from their home campus. Students can also borrow books and other circulating material as well as access full-text journal articles.

General Library Support Links:

- Global Campus
- Pullman Campus
- Spokane Campus
- Tri-Cities Campus
- Vancouver Campus
- College of Nursing

Online Tutoring

As a WSU student enrolled in an undergraduate course, you have free unlimited access to Online Tutoring. This is not a course requirement, but a resource for you to utilize as needed.

With three ways to access a tutor you can choose the one that best fits your needs:

- Submit a paper: Writing Lab tutors will respond to papers in any academic subject. Just submit your paper, ask specific questions on the submission form, and a tutor will respond within 24-48 hours.
- Live tutoring: eChat rooms allow students to meet with tutors in one-on-one tutoring sessions via a fully interactive, virtual online environment.
- Leave a question: Students can leave specific questions for a tutor in any of our subjects by taking advantage of our eQuestions option. Our tutors will respond to your question within 24-48 hours.

More details and the list of available tutoring subjects can be found at www.eTutoringOnline.org

WSU Academic Calendar

Please refer to the WSU academic calendar to be aware of university holidays and important deadlines throughout the semester.

Copyright

Any course-related materials, presentations, lectures, etc. are the instructor's intellectual property and may be protected by copyright. The use of University electronic resources for commercial purposes, including advertising to other students to buy notes, is a violation of WSU's computer abuses and theft policy (WAC 504-26-218). Selling class notes through commercial note taking services without written advance permission from the faculty, could be viewed as be as copyright infringement and/or academic integrity violation(WAC 504-26-010 (3)(a,b,c,i)).

University Syllabus

Students are responsible for reading and understanding all university-wide policies and resources pertaining to all courses (for instance: accommodations, care resources, policies on discrimination or harassment), which can be found in the University Syllabus.